Student Learning Outcomes and Attitudes Using Three Methods of Group Formation in a Nonmajors Biology Class

Deborah A. Donovan,[†]* Georgianne L. Connell,[†] and Daniel Z. Grunspan[§]

[†]Biology Department and [‡]Science Education Group, Western Washington University, Bellingham, WA 98225; [§]Center for Evolution and Medicine, Arizona State University, Tempe, AZ 85281

ABSTRACT

Group work is often a key component of student-centered pedagogies, but there is conflicting evidence about what types of groups provide the most benefit for undergraduate students. We investigated student learning outcomes and attitudes toward working in groups when students were assigned to groups using different methods in a large-enrollment, student-centered class. We were particularly interested in how students entering the class with different levels of competence in biology performed in homogeneous or heterogeneous groups, and what types of group compositions were formed using different methods of group formation. We found that low-competence students had higher learning outcomes when they were in heterogeneous groups, while mid- and high-competence students performed equally well in both group types. Students of all competence types had better attitudes toward group work in heterogeneous groups. The use of student demographic variables to preemptively form groups and allowing students to self-select their group mates both yielded heterogeneous competence groups. Students in the instructor-formed, demographic groups had higher learning outcomes compared with students allowed to self-select. Thus, heterogeneous groupings provided the most benefit for students in our nonmajors, large-enrollment class.

INTRODUCTION

Group work is often a key component of student-centered pedagogies, in which students, rather than the instructor, are the focus of instruction and students are frequently engaged in activities that promote higher-order thinking. Productive group dynamics are important for peer instruction (Mazur, 1996), team-based learning (Michaelsen *et al.*, 2004), cooperative learning (Johnson *et al.*, 1991), and course-based undergraduate research experiences (CUREs) (Dolan *et al.*, 2008; Corwin *et al.*, 2015), as well as for active-learning strategies like think–pair–share, jigsaw, reciprocal teaching, and case studies. The importance of understanding what makes a productive group increases as the push to increase the use of active-learning pedagogies across college classrooms grows (Handelsman *et al.*, 2004; Ruiz-Primo *et al.*, 2011; Freeman *et al.*, 2014).

Two general categories of group work exist in classroom settings, informal and formal (Smith and McGregor, 1992; Tanner *et al.*, 2003). Informal groups are formed for a short period of time, such as when students are asked to discuss a question or concept with a neighbor during think–pair–share. Formal groups are more fixed and longer lasting, such as when students work in groups throughout the duration of a course on structured classroom activities. The latter can be particularly powerful when students work together in cooperative learning groups, in which students work toward shared learning goals and are assessed on both group and individual work (Johnson *et al.*, 1991; Tanner *et al.*, 2003).

Cooperative learning is grounded in social interdependence theory, which posits that learning outcomes of students working in groups are affected by the nature of the

relationships within the group (Johnson and Johnson, 2009; Johnson *et al.*, 2014). Learning is enhanced when there is positive interdependence, in which a student's success is linked to the success of others in the group. Johnson *et al.* (2014) found that, when interdependence is positive, students tend to engage in promotive interactions, which are characterized by behaviors such as helping, sharing, and encouraging one another. When interdependence is negative, such as in competitive groups, a student's achievements are negatively correlated with those of his or her group. No interdependence exists when a student's achievements are not linked with others' achievements, such as when students are working individually (Johnson *et al.*, 2014).

Working in groups has been linked to increased student achievement and better student attitudes in college classrooms. Springer *et al.* (1999) analyzed studies on small-group learning in undergraduate science, math, engineering, and technology classes. They found that students working in groups had higher academic achievement, better attitudes toward learning, and increased persistence in class work compared with students in more traditional classes that lacked group work. A more recent analysis of higher education studies from multiple subject areas and different countries found that students working in cooperative groups had higher individual achievement compared with students working in competitive groups with negative interdependence or in an individualistic environment with no interdependence (Johnson *et al.*, 2014). Smith *et al.* (2011) also found that cooperative types of activities, such as engaging in peer discussion while answering clicker questions, increased a student's ability to correctly answer related questions compared with students who only heard an instructor's explanation about the question.

While there is consensus that cooperative learning is effective, there is less agreement about how to structure groups to maximize learning. Many instructors follow the recommendations for team-based learning, in which groups are structured to include students with diverse perspectives, backgrounds, and academic characteristics (see Michaelsen and Sweet, 2008). Heterogeneity in groups can be based on many different student characteristics: both academic (e.g., prior course work) and personal (e.g., demographic). For example, McInerney and Fink (2003) found that student achievement increased in an undergraduate microbial physiology course when it transitioned to using team-based learning with students heterogeneously grouped based on previous experience in microbiology and chemistry. Using academic major as a basis to sort students into heterogeneous groups, Gaudet *et al.* (2010) transformed an upper-level neurobiology course from an individualistic structure to a small-group structure. They found that lower-performing students did significantly better when they were working in groups, compared with individually. They also found that high-performing students did better on group quizzes compared with individual quizzes, indicating that high-performing students can also benefit from group work.

It is also important to consider student demographics when forming heterogeneous groups. Although it may seem appropriate to evenly distribute minority students (either ethnic or gender) among groups, Rosser (1998) describes the importance of not isolating minority students when forming groups. Indeed, when students are allowed to self-select into groups, they tend

to choose group members of the same gender and ethnicity (Freeman *et al.*, 2017).

Other important considerations regarding group formation include the academic ability and content knowledge of each student in the group. In this context, ability can mean many things, ranging from self-regulated learning, previous exposure to the material, or any other factor that contributes to success in a class, while content knowledge refers to general familiarity with the class material. Student groups inevitably differ in their composition of student ability and knowledge, and thus differ in degrees of heterogeneity.

To date, most research on the impacts of group composition with respect to ability has been done at the K–12 level (e.g., Kulik and Kulik, 1982, 1984; Slavin, 1990), though there is some research from postsecondary science classrooms. Table 1 summarizes studies from a range of university courses. Of the experimental courses, all but one were solely at the undergraduate level (one course included graduate students). The meta-analyses included studies of K–12 through postsecondary classrooms. It is clear from these studies that there is no consensus about whether homogeneous or heterogeneous groups (based on ability) are better for students. Furthermore, in studies that compared outcomes for students of different abilities, the results were often different for low-, mid-, and high-performing students.

Another common method of forming groups is to allow students to self-select into their groups. This method requires little planning for the instructor and may yield groups that work together better from the start, because students often select group mates they already know (Strong and Anderson, 1990; Bacon *et al.*, 1999). However, there are drawbacks in allowing students to self-select their groups. In investigating why groups fail, Feichtner and Davis (1984) found that students self-reported by a 2 to 1 margin that their worst group experiences were when groups were self-selected, although Bacon *et al.* (1999) found that more students reported that their best experiences were in classes in which they were in control of forming their own teams.

The evidence is mixed about the effects of self-selected groups on student achievement compared with instructor-formed groups. For example, undergraduates allowed to self-select into laboratory groups had lower posttest scores compared with students in instructor-formed groups (both heterogeneous and homogeneous), although there were no differences between students in the heterogeneous and homogeneous groups (Lawrenz and Munch, 1984). Similarly, Brickell *et al.* (1994) found that engineering students allowed to self-select into groups had lower group grades and poorer attitudes about the class compared with students who had been assigned to groups. However, Theobald *et al.* (2017) found that students who were more comfortable in their group had better test scores than students who were not comfortable and that having a friend in the group was the best predictor of comfort. Interestingly, there was no effect of having a friend in the group on student performance. One concern about self-selected groups is that they often lack diversity of student skills and perspectives (Mello, 1993; Bacon *et al.*, 1998). In this respect, Freeman *et al.* (2017) found that, when students were allowed to self-select into groups in a large-enrollment biology class, students tended to work with peers of the same ethnicity and gender and similar academic ability.

TABLE 1. Results of studies conducted in college and university courses in which students were grouped by academic ability into homogeneous or heterogeneous groups

Group type that most benefited students	Context and results	Reference
No difference	Undergraduate physical science course for preservice elementary teachers. There was no difference between posttest scores of students grouped by performance on a reasoning assessment.	Lawrenz and Munch (1984)
Heterogeneous	Undergraduate physics course. Problem-solving ability of students in heterogeneous groups was better than students in homogeneous groups.	Heller and Hollabaugh (1991)
Homogeneous (weak support)	Introductory, undergraduate, life science course. Students in homogeneous groups performed slightly, but not significantly, better than students in heterogeneous groups.	Watson and Marshall (1995)
Homogeneous	Meta-analysis of studies from elementary through postsecondary classrooms. Homogeneous groups were better overall. Mid-ability students learned more in homogeneous groups; low-ability students learned more in heterogeneous groups; no difference for high-ability students	Lou et al. (1996)
Homogeneous (weak support)	Meta-analysis of studies from elementary through postsecondary classrooms. Results generally supported Lou et al. (1996), although group type was not a significant predictor in the models.	Lou et al. (2000)
Homogeneous	Undergraduate psychology course. Mid- and high-achieving students learned more in homogeneous groups; no difference for low-achieving students.	Baer (2003)
Homogeneous	Introductory, undergraduate, life science course. Low-reasoners in homogeneous inquiry groups outperformed low-reasoners in heterogeneous groups; no differences for mid- and high-reasoners.	Jensen and Lawson (2011)
Heterogeneous	Upper-level biotechnology lab. Students paired with a student of a different academic level (undergrad and grad) earned better grades than students paired with another student at the same level.	Miller et al. (2012)
No difference	Undergraduate physics students. No differences between students working in homogeneous or heterogeneous groups.	Harlow et al. (2016)

The increased use of formal groups in undergraduate biology classes, coupled with conflicting evidence about how to best form groups to maximize student learning, led us to investigate student learning outcomes and attitudes toward working in groups when students were assigned to groups using different methods. We were particularly interested in how students entering the class with different levels of competence in biology performed in homogeneous or heterogeneous groups, and what types of group compositions were formed using different methods of group formation. In the first experiment, we used performance on a preassessment to identify the competence level of students as they entered the class, then assigned students to either heterogeneous or homogeneous competence groups. In a subsequent class, we assigned students to heterogeneous groups based on grade point average (GPA) and number of science classes completed in lieu of a time-consuming preassessment, then determined the types of competence groups that had formed and measured performance outcomes. Finally, we allowed students to self-select into groups and determined the types of competence groups that formed. In all classes, we measured student learning outcomes and attitudes toward working in groups.

We hypothesized that:

1. Student learning outcomes of high-, mid-, and low-competence students would differ between heterogeneous and homogeneous competence groups.
2. Student attitudes toward working in groups of high-, mid-, and low-competence students would differ between heterogeneous and homogeneous competence groups.
3. Heterogeneous competence groups could be assigned without relying on a preassessment by instead using student demographic variables and measures of prior academic achievement that are predictive of preassessment score.
4. Allowing students to self-select into their groups would result in a greater number of homogeneous competence groups than expected by chance.
5. Student learning outcomes and attitudes toward working in groups would differ between classes in which the instructor assigned students into groups compared with those in which student groups were self-selected.

METHODS

Our study was approved by the Human Subjects Review Committee at Western Washington University (IRB# FWA00001207).

Classroom Context

We conducted our study across three iterations of a large-enrollment, nonmajors, student-centered biology class at a regional, primarily undergraduate institution. At our institution, Biology 101 is taught in sections of ~200 students in a large lecture hall with fixed seating. Groups of four to six students are formed on the first day of class, and students remain in their groups for the duration of the quarter. The class has been "flipped" for approximately 10 years, and content is covered in a series of 1- to 2-week-long modules. At the start of a module, students individually complete online work outside class, which includes watching an online lecture, taking an online quiz, and posting areas of confusion to a discussion board. When students are in class, they work in their groups to complete activities such as worksheets, jigsaws, group quizzes, and group exams (they also take each exam individually before taking it as a group). Thus,
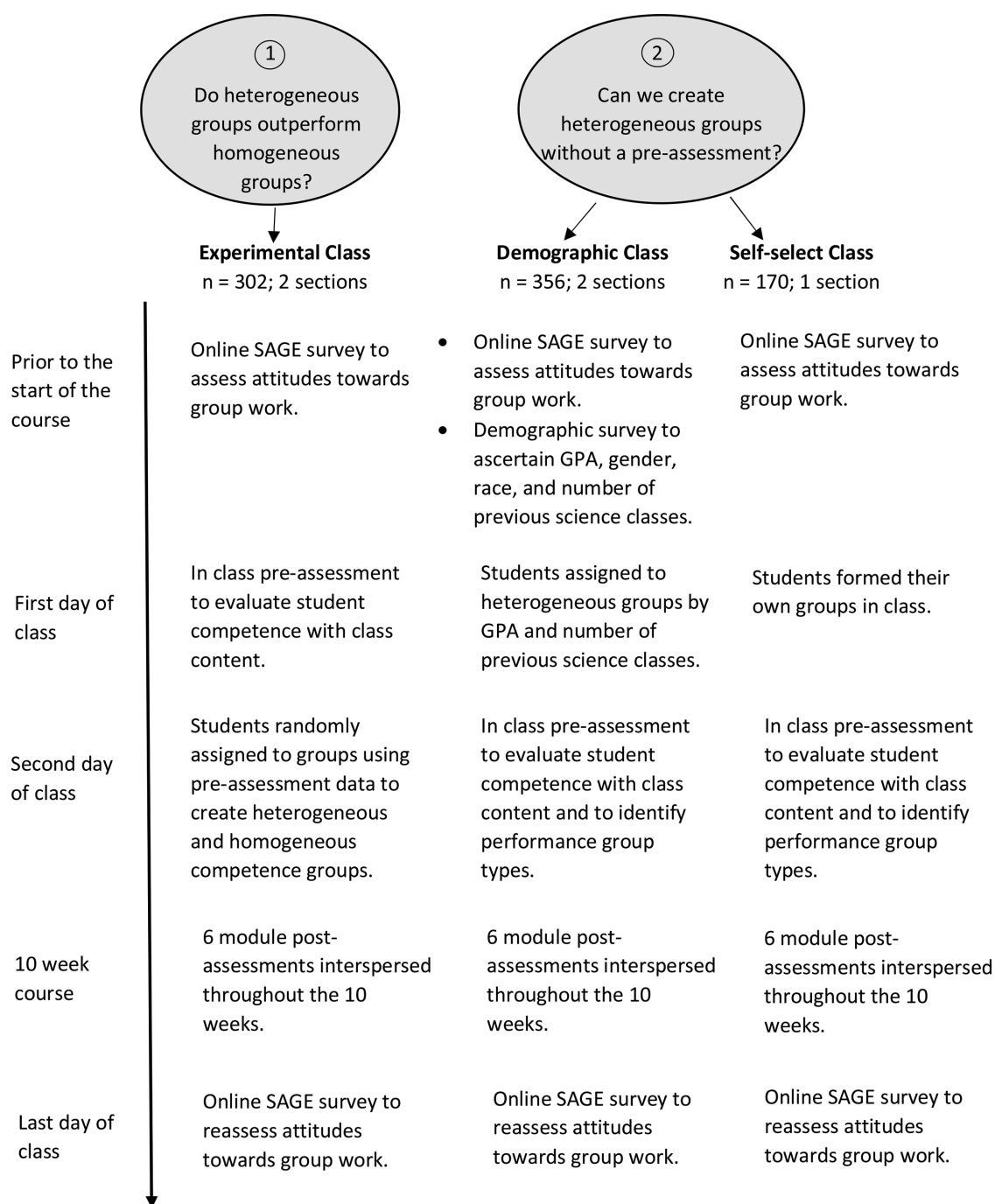
FIGURE 1. Structure of the three classes in this study, including types of assessments and how they were administered.

each student works closely with his or her group members throughout the class. Student work is organized with folders, which allows the instructor to pass out and collect group work, organize group quizzes and exams, and communicate with the groups. See Connell *et al.* (2016) for a more thorough description of the class content and structure.

In the present study, we experimented with three different ways to form groups and measured learning outcomes and students' attitudes toward working in groups (Figure 1). The same instructor (G.L.C.) taught all classes, using the same materials and the same assessments. The study took place during three quarters over the span of 1.5 years.

Content and Attitude Assessments

We used the same content and attitude assessments in each class to assess learning outcomes and attitudes about working in groups. We developed the content assessment using modified versions of concept inventory questions (Klymkowsky and

Garvin-Doxas, 2008; D'Avanzo *et al.*, 2010; Nadelson and Southerland, 2010; Fischer *et al.*, 2011) and questions that we wrote ourselves (available in the Supplemental Material). The assessment consisted of 60 multiple-choice questions developed to cover the range of material in the class. The same content assessment was administered as a preassessment and as a postassessment. The preassessment was taken individually on the first day of class. The postassessment was administered in sets of questions ("exams") after a module (i.e., the cell biology and genetics questions were administered after the cell biology and genetics modules).

Attitudes toward working in groups were assessed using the Student Attitudes towards Group Environments (SAGE) survey (Kouros and Abrami, 2006). The SAGE is a five-point Likert survey that probes student attitudes about four factors of working in groups: quality of product and process (i.e., "When I work in a group I do better quality work"), peer support (i.e., "When I work in a group I am able to share my ideas"), student interdependence (i.e., "Everyone's ideas are needed if we are going to be successful"), and frustration with group members (i.e., "I become frustrated when my group members do not understand the material"). The SAGE is composed of 43 questions, with 8–15 questions per factor. Several of the questions are negatively worded on the survey and were reverse coded during analysis following the methods of Kouros and Abrami (2006). All of the questions in the frustration with group members factor are negatively worded, so, after reverse coding, the factor as we report it more accurately represents "satisfaction with group members." We administered the SAGE through our course management system before the first day of class and again at the end of the quarter.

Learning Outcomes and Attitudes in the Experimental Class of Students in Heterogeneous and Homogeneous Competence Groups (Hypotheses 1 and 2)

In our first experiment, we gave students ($n = 302$ in two sections of Biol 101) the preassessment to determine their incoming competence in biology, then assigned students to different types of competence-based groups (we refer to this class as Experimental). We used the term "competence" in lieu of "ability," because our preassessment was a single measure of students' incoming competence with biology as opposed to a more general ability to learn biology. Students in the Experimental class were given the preassessment on the first day of class, and it was used to rank them as high-pretest score (HPS) students, mid-pretest score (MPS) students, or low-pretest score (LPS) students. To assign the rankings, we created a histogram of all preassessment scores and looked for breaks in the scores. After this ranking, we had 65 HPS students, 168 MPS students, and 69 LPS students. We then used a random number generator to place students into either heterogeneous or homogeneous groups (Table 2). Heterogeneous groups had at least one HPS student, one MPS student, and one LPS student in each group of four to six students. Homogeneous groups had all HPS, all MPS, or all LPS students. After assigning groups, we had 32 heterogeneous groups, seven low-homogeneous groups, 19 mid-homogeneous groups, and six high-homogeneous groups. Students sat in assigned seats every day, and neither the students nor the instructor was aware of the types of groups that had been formed.

Analyses for Hypotheses 1 and 2

Our observations are on students, but students are clustered into groups. When data have these properties, the observations (and the errors associated with them) are not independent. Multilevel models (MLMs; also known as hierarchical models or mixed models) effectively handle this nonindependence by distinguishing between within-level variation and between-level variation (Raudenbush and Bryk, 2002; Goldstein, 2011; Theobald, 2018). MLMs are able to model effects at the lowest level (in this case students) by modeling the variation in the higher-level variables (in this case the groups). These higher-level variables (e.g., groups) are the variables that make the lower-level variables (e.g., students) nonindependent. Thus, we used MLMs to control for issues with nonindependence by

TABLE 2. The types of groups formed in this study and the classes in which they occurred

Class	Group types (% of groups in class)	Student composition
Experimental	Homogeneous groups	
	Low (11)	All LPS
	Mid (30)	All MPS
	High (9)	All HPS
	Heterogeneous groups	
	Low–Mid–High (50)	LPS, MPS, HPS
Demographic	Homogeneous groups	
	Mid (1)	All MPS
	Heterogeneous groups	
	Low–Mid–High (58)	LPS, MPS, HPS
	Low–Mid (33)	LPS, MPS
	Mid–High (8)	MPS, HPS
Self-selected	Homogeneous groups	
	Mid (3)	All MPS
	Heterogeneous groups	
	Low–Mid–High (66)	LPS, MPS, HPS
	Low–Mid (22)	LPS, MPS
	Mid–High (9)	MPS, HPS

distinguishing the intragroup variation and model effects of variables at the student level. These models were estimated using the *lmer* function, with the *lme4* package in R (Bates *et al.*, 2015).

We were specifically interested in whether learning gains of individual students were affected by the composition of their groups, and whether any potential effect depended on the students' competence. Using data from the Experimental class, we fit two-level MLMs with individual student post scores as the dependent variable, and a random effect for student group to capture variation in scores that may be attributable to unique aspects of each group. To test whether student learning outcomes differed between LPS, MPS, and HPS students in homogeneous or heterogeneous groups, we developed and fit a series of four alternative models. We then compared corrected Akaike information criterion (AICc) scores, which adds a correction for small sample sizes, to test for the best-fitting model, with ΔAICc ≤ 2 between models set as a threshold for equivalent fit (Anderson and Burnham, 2002). The fixed effects in the four alternative models were 1) a null model with no fixed effects; 2) only pretest score; 3) pretest score and group type (heterogeneous or homogeneous); and 4) pretest score, group type, and an interaction between group type and student competence (LPS, MPS, or HPS). It is important to realize that each of these models is an alternative hypothesis, and using model selection is the way to test between them.

To examine whether group composition affected student attitudes about group work, we performed similar MLM analyses as those described earlier with student SAGE responses as the dependent variables of interest. Student SAGE responses from the beginning of class and then again at the end of class were averaged for each component of the survey, providing pre and post scores between 1 and 5 for quality of product and process, peer support, student interdependence, and frustration with group members. There was no evidence of a ceiling effect for the SAGE scores (Supplemental Figure 1), which can be common for Likert survey data. To understand whether individual student attitudes changed as a function of group type and student competence, we tested a set of five alternative models for each of the four SAGE scores and used the AICc criteria described earlier for model selection. The fixed effects in the five alternative models were 1) a null model with no fixed effects; 2) only the pretest score of the SAGE factor of interest (pretest SAGE score); 3) pretest SAGE score and group type; 4) pretest SAGE score, group type, and student competence; and 5) pretest SAGE score, group type, student competence, and an interaction between group type and student competence.

Types of Competence Groups Formed Using Demographic Variables and Measures of Prior Academic Achievement (Hypothesis 3)

As described in the *Results*, LPS students were most successful when working with higher-competency students. For this reason, we sought to replicate heterogeneous group formation in the quarter following the Experimental study by using demographic variables and measures of prior academic achievement instead of a time-consuming in-class pretest. In this class (*n* = 356 in two sections of Biol 101), we collected demographic and achievement data in a preclass survey and used those data to proactively assign students to create heterogeneous groups. We

subsequently measured each student's preassessment score on the first day of class to determine the competence composition of each group (we refer to this class as Demographic).

The preclass survey was administered before the first day of class via our course management system and asked for information for ~10 variables: self-reported college GPA, self-rating of proficiency in biology (novice, competent, proficient), number of other science classes taken at the college level, years of high school biology, year in university, age, first-generation college student status, comfort with the English language (very uncomfortable, uncomfortable, comfortable, very comfortable), gender, and race/ethnicity. To assign groups, we used two variables that we thought would best predict preassessment score: self-reported GPA, because it can predict success in introductory biology courses (Freeman *et al.*, 2007), and number of previous science classes taken at our university, because it predicted course success in a previous study at our institution (Connell *et al.*, 2016). We also used gender and race/ethnicity to ensure students had allies in their groups following the recommendations of Rosser (1998). On the first day of class, students completed the preassessment, and we used their scores to rank them as HPS, MPS, or LPS students so we could determine the composition of the groups that had been formed. We used the same breaks in preassessment scores that we used in the Experimental class to categorize LPS, MPS, and HPS students.

To determine whether the other variables in our survey were predictors of preassessment score, we performed a post hoc multiple linear regression using nine of the variables collected in the demographic survey (self-reported college GPA, self-reported proficiency in biology, the number of science classes a student took at the college or university level, the number of high school biology classes taken, the student's year in college, first-generation student status, language proficiency, and gender). We did not include race/ethnicity in the regression, because the number of ethnic minorities was very low in our class. We performed model selection among all possible combinations from this global model using AICc and the model-selection criteria previously described to determine which variables were most predictive of prescore tests. This model selection was performed using the *dredge* function in the *MuMIn* package in R (Barton, 2009).

Types of Competence Groups Formed When Students Self-Select into Their Groups (Hypothesis 4)

Allowing students to self-select into their groups is another common method of forming groups. To examine outcomes of this method, we allowed students (*n* = 170 in one section of Biol 101) to select their own groups. Before the first day of class, we alerted students that they would be working in groups and that they would organize themselves into groups on the first day of class (we refer to this as the Self-selected class). In class, students were directed to move into groups. The instructor helped the 10 or so students who could not find a group on their own. Students then completed the preassessment in class, and we used their scores to rank them as HPS, MPS, or LPS students, using the same breaks in scores as in the other classes of our study, to determine the types of competence-based groups that had been formed when students selected their own groups.

To understand group composition in this classroom, we compared the realized proportion of different group types to the

average proportion of group compositions from 1000 simulated classes of the same size (n = 170) forming randomized groups of five students (the average group size in the actual class was 5.3). The resulting frequencies of groups formed allow an understanding of what kinds of groups students may be more or less likely to form than would be expected if students mixed completely at random.

Student Learning Outcomes and Attitudes toward Working in Groups in Classes with Self-Selected and Instructor-Formed Groups (Hypothesis 5)

To test whether the assignment of students into groups by the instructor had any impact on student experiences, we examined student post scores on the assessment and SAGE results of students in the Demographic (instructor-formed) and Self-selected classes. We excluded the Experimental class from this analysis, because its distribution of group types was much different from the other two classes due to the deliberate formation of homogeneous and heterogeneous groups.

To test whether learning outcomes differed between students in these classes, we tested two alternative MLMs with post score as the dependent variable and student group modeled as a random variable. Both models included student GPA (obtained from the registrar), pretest scores, and group composition as controls for expected student performance, while only the second model also included a fixed variable for method of group formation (demographic or self-selection). While it would typically be ideal to model class section as a random variable, this is not possible, because the treatment (demographic or self-selection) is defined by class section. We compared AICc scores of the full model and a reduced model without the class variable.

A second set of models were run for each SAGE factor postclass score, with class as the main independent variable of interest and SAGE prescore, performance on the class pretest, group type, and student GPA as control variables. As for testing learning gains, we compared the AICc scores of a full model including class with a reduced model excluding class to test whether student group dynamics differed between these classes.

RESULTS

Almost all students experienced an increase in performance on the postassessment, with only three students experiencing zero change in score and two students scoring lower on the postassessment compared with the preassessment. Scores on the postassessment ranged from 15 to 58 points. Changes in

SAGE scores were generally positive, except for interdependence which decreased on average in the Demographic and Self-selected classes. Scores for quality of product, peer support, and frustration (satisfaction) generally increased by 0.1 to 0.5 points on the Likert scale between the beginning and end of the class. Means and standard deviations of pre- and postassessment scores (out of 60 possible points) and pre- and post-SAGE scores in each of the three different classes are shown in Table 3.

Learning Outcomes and Attitudes of Students in Heterogeneous and Homogeneous Competence Groups (Hypotheses 1 and 2)

Group type was a significant predictor of performance on the posttest for LPS students; they had higher learning gains when in heterogeneous groups than when in groups with only other LPS students. According to our model-selection criteria, the full model, which included an interaction between group type and student performance, was the best-fitting model for student posttest scores (Table 4). The size of the gain for LPS students in heterogeneous groups compared with those in homogeneous groups was an estimated 3.2 points, equivalent to 5.3% of the possible points on the 60-point assessment test. The interaction term indicates that neither MPS students nor HPS students experienced the same learning gains from heterogeneous groups that LPS students did. MPS students scored an estimated 0.697 points (3.218 − 3.915 = −0.697) worse on average when in heterogeneous groups compared with homogeneous groups, while HPS students scored an estimated 1.085 points (3.218−4.303) worse on average in heterogeneous groups. Figure 2 displays the raw data with model estimates from the MLM to provide a qualitative understanding of the associations between student competency, group composition, and post scores, and the best fit of the model.

For the SAGE scores, the best-fitting models for students' frustration (satisfaction) with group members and their reported quality of product and process included the variable for group type. In both constructs, students reported significantly higher affect in heterogeneous groups compared with students in homogeneous groups, independent of student competence, indicating better group work in heterogeneous groups for these constructs compared with homogeneous groups (Figure 3 and Supplemental Table 1). There was no indication that reported student interdependence or peer support was associated with either group composition or student competence. In fact, student performance on the content preassessment had no association

TABLE 3. Summary statistics from each of the three classes[a]

	Class					
	Experimental		Demographic		Self-selected	
	Pre	Post	Pre	Post	Pre	Post
Content assessment	21.45 ± 5.16	38.70 ± 7.59	20.27 ± 5.36	39.71 ± 7.99	21.15 ± 5.54	37.03 ± 8.52
SAGE factors						
Quality	3.29 ± 0.65	3.58 ± 0.69	3.45 ± 0.53	3.54 ± 0.58	3.34 ± 0.68	3.52 ± 0.67
Interdependence	3.85 ± 0.41	3.83 ± 0.53	3.88 ± 0.39	3.63 ± 0.46	3.85 ± 0.43	3.65 ± 0.53
Peer support	3.74 ± 0.47	3.82 ± 0.56	3.77 ± 0.43	3.89 ± 0.50	3.74 ± 0.47	3.81 ± 0.59
Frustration (satisfaction)	2.92 ± 0.50	3.34 ± 0.56	3.04 ± 0.44	3.26 ± 0.54	2.93 ± 0.48	3.13 ± 0.53

[a]Pre and post content assessment scores were out of 60 total points. Each SAGE factor (quality of product and process, interdependence, peer support, and frustration) was on a five-point scale. Numbers represent means ± SD.

TABLE 4. Summary of fixed effects from multilevel regression analyses for variables predicting postassessment scores of students in the Experimental class[a]

Effect	Model 1		Model 2		Model 3		Model 4	
	Estimate ± SE	*p* value	Estimate ± SE	*p* value	Estimate ± SE	*p* value	Estimate ± SE	*p* value
Intercept	**38.961 ± 0.538**	**<0.001**	**22.439 ± 1.622**	**<0.001**	**22.387 ± 1.674**	**<0.001**	**19.597 ± 2.183**	**<0.001**
Pretest score			**0.764 ± 0.073**	**<0.001**	**0.764 ± 0.073**	**<0.001**	**0.893 ± 0.098**	**<0.001**
Group type (ref: homogeneous)								
Heterogeneous					0.126 ± 0.809	0.876	**3.218 ± 1.393**	**0.022**
Group type × performance (ref: LPS)								
Heterogeneous × MPS							**−3.915 ± 1.426**	**0.006**
Heterogeneous × HPS							**−4.303 ± 2.048**	**0.037**
AICc	2084.2		1998.8		1999.4		1990.8	
ΔAICc	93.4		8		8.6		—	

Statistically significant estimates for each model are in bold text.

[a]Low-competency (LPS) students scored between 0 and 17 points on the pretest, mid-competency (MPS) students scored between 18 and 25 points on the pretest, and high-competency (HPS) students scored between 26 and 39 points on the pretest. These analyses support hypothesis 1.

with any of the SAGE constructs. Thus, we found no evidence that LPS, MPS, or HPS students differed in their attitudes toward working in groups, even in homogeneous groups, in which LPS student groups performed particularly poorly compared with MPS and HPS students.

Types of Competence Groups Formed Using Demographic Variables and Measures of Prior Academic Achievement (Hypothesis 3)

The types of groups formed in the Demographic class differed from the types in the Experimental class (Table 2). Only one of the 64 groups was homogeneous, and it was a mid-homogeneous group. The other 63 were heterogeneous, but many were different in terms of composition compared with the Experimental class, in which heterogeneous groups always consisted of at least one HPS, one MPS, and one LPS student. In the Demographic class, 37 of the groups were heterogeneous with at least one HPS, one MPS, and one LPS student, similar to the Experimental class. Twenty-one had only MPS and LPS students and five had only HPS and MPS students. There were no groups with only HPS or LPS students. Thus, using demographic variables and measures of prior academic achievement resulted in many types of heterogeneous groups, and almost no homogeneous groups, as was the goal.

When we tested how well student variables predicted their preassessment scores, we found that the full model with all
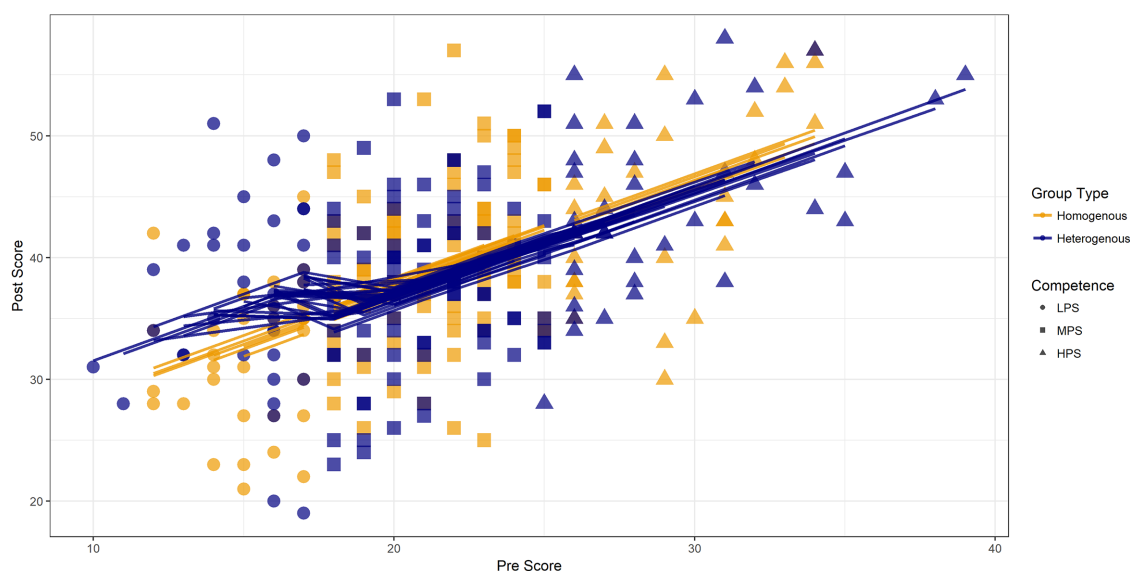

FIGURE 2. Individual postassessement scores by preassessment performance and group type, with lines indicating model estimates from the MLM for each student group. Low-competence (LPS) students in heterogeneous groups performed better on average compared with LPS students in homogeneous groups, as indicated by the separation between model estimates for homogeneous groups and heterogeneous groups for students with lower preassessment scores. The difference in postassessment scores by group type for mid-competence (MPS) and high-competence (HPS) students is in the opposite direction, but is much smaller. This analysis supports hypothesis 1.
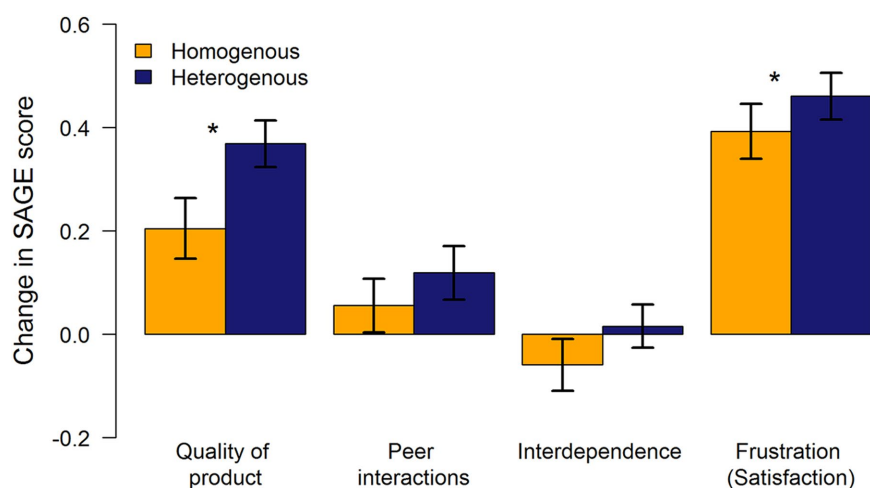
FIGURE 3. Change in the four SAGE factors of low-, mid-, and high-competence (LPS, MPS, and HPS, respectively) students in heterogeneous and homogeneous groups. Error bars represent SE. Change in quality of product and frustration (satisfaction) was significantly greater for students in heterogeneous groups compared with students in homogeneous groups. This analysis supports hypothesis 2.

proficiency in biology. When we compared models describing all possible combinations of the demographic variables, 40 models had ΔAICc $\leq$ 2, indicating they had the same predictive value. Only two variables, self-reported GPA and self-rating in biology, were included in these 40 models, and the model that contained only those two variables was among the top 10 best-fitting models according to AICc (Supplemental Table 2). Thus, we selected the model containing only self-reported GPA and self-rating in biology as the most parsimonious, following the recommendations of Burnham and Anderson (2002).

Types of Competence Groups Formed When Students Self-Select into Their Groups (Hypothesis 4)

We examined how students in the Self-selected class assorted into groups by simulating what groups would be expected if students were to assort completely at random and comparing the simulation results to the groups that formed in this class. Figure 4 shows the expectations from 1000 simulations against the proportions of groups that were realized in the Self-selected class. The proportions of group types from the Demographic class have been included in the figure for additional comparison. While the group compositions were not strikingly

variables explained only 16% of the variation in preassessment score (Table 5). Of the two variables used to assign groups in our study (self-reported GPA and number of previous science classes taken at college or university level), only self-reported GPA significantly predicted preassessment score. The only other significant variable from our demographic survey was self-rating of

TABLE 5. Estimated regression coefficients from a multiple linear regression used to determine whether a student's preassessment score was predicted by different demographic variables

Regression coefficients	Estimate ± SE	*p* value
Model intercept (β_0)	9.19 ± 3.15	0.004
Self-reported GPA (β_1)	**1.48 ± 0.27**	**<0.001**
Self-rating in biology (reference level: novice) (β_2)		
Competent	**1.35 ± 0.57**	**0.018**
Proficient	**3.07 ± 1.18**	**0.009**
Number of other science courses (β_3)	−0.15 ± 0.28	0.59
Years of high school biology (reference level: none) (β_4)		
One	1.68 ± 1.24	0.17
Two (AP Biology)	2.20 ± 1.72	0.20
Year in university (reference level: freshman) (β_5)		
Sophomore	−0.25 ± 0.65	0.70
Junior	−0.72 ± 0.94	0.45
Senior	−1.28 ± 1.31	0.33
Age (β_6)	0.10 ± 0.14	0.48
First-generation student (reference level: no) (β_7)		
Yes	0.71 ± 0.61	0.24
Comfort with English language (reference level: comfortable) (β_8)		
Very comfortable	1.46 ± 1.08	0.18
Gender (reference level: female)		
Male	0.47 ± 0.58	0.42
Other	5.43 ± 3.03	0.07

Statistically significant estimates for each model are in bold text.

The r^2 for the full model regression equation was 0.16. The *p* values are the results of *t* tests to determine whether the slope (β) of each variable was significantly different from 0.
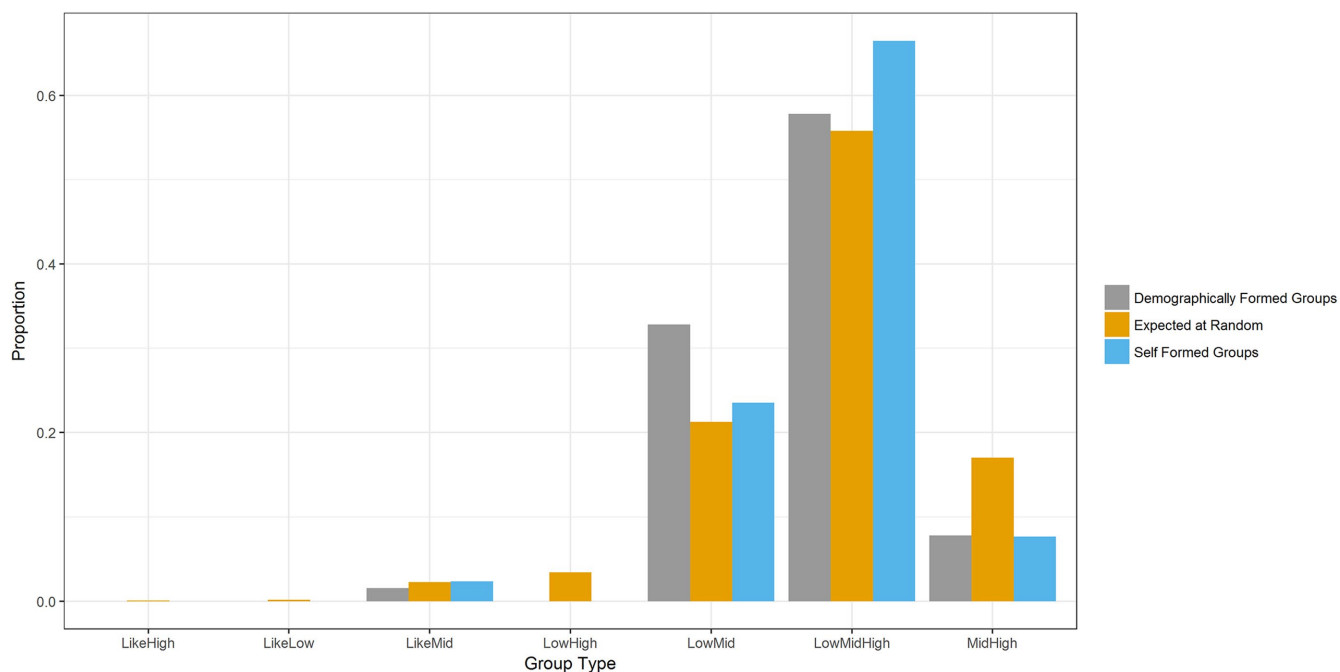
FIGURE 4. Observed group types realized when students were allowed to self-select their groups compared with group types predicted by random assortment, which was determined by 1000 simulations. Error bars for the simulated groups, representing SD, are present but very small. This analysis supports hypothesis 4.

different from what we may expect at random, it is noteworthy that students self-selected heterogeneous groups at a slightly higher rate than expected by chance, and the only homogeneous group type formed in both the self-selected and demographically formed classes were all MPS students. The heterogeneous groups that formed were similar between both the Self-selected class and the Demographic class (Table 2 and Figure 4).

Student Learning Outcomes and Attitudes toward Working in Groups in Classes with Self-Selected and Instructor-Formed Groups (Hypothesis 5)

Students in the Self-selected class scored significantly lower on the posttest than students in the Demographic class after controlling for group type, student performance on the pretest, and student GPA (obtained from the registrar) (Table 6). The

best-fitting model according to our model-selection criteria was the full model that included course. However, this result must be treated with caution, because the classes were taught during different quarters, even though they were taught by the same instructor using the same materials.

Despite the difference in learning outcomes between these two classes, including class in our models for the four SAGE constructs did not improve model fit, indicating no significant difference in student attitudes toward working in groups (Supplemental Table 3).

DISCUSSION

Heterogeneous versus Homogeneous Competence Groups
In the Experimental class, in which heterogeneous and homogeneous groups were intentionally formed based on

TABLE 6. Summary of fixed effects from MLM analyses for variables predicting posttest scores between the Demographic and Self-selected classes[a]

Effect	Model 1		Model 2	
	Estimate ± SE	*p* value	Estimate ± SE	*p* value
Intercept	**0.439 ± 1.875**	**<0.001**	**12.010 ± 1.866**	**<0.001**
GPA	**5.047 ± 0.519**	**<0.001**	**5.153 ± 0.514**	**<0.001**
Pretest score	**0.588 ± 0.055**	**<0.001**	**0.607 ± 0.054**	**<0.001**
Group composition (reference: homogeneous)				
Heterogeneous	**2.842 ± 1.037**	**0.007**	0.446 ± 1.053	0.666
Section (reference: Demographic)				
Self-selected			**−3.625 ± 0.735**	**<0.001**
AICc	3483.7		3462.4	
ΔAICc	21.3		—	

Statistically significant estimates for each model are in bold text.
[a]These analyses support hypothesis 5.

student competence in biology, the lowest-performing students experienced higher learning outcomes, higher group satisfaction, and greater perception of the quality of their work when placed in heterogeneous groups compared with working with peers of similar competence. At the same time, differences in learning outcomes were negligible for high- and mid-performing students between heterogeneous and homogeneous groups. Our result for LPS students corroborates meta-analyses by Lou *et al.* (1996, 2000) who found that low-ability students benefited from heterogeneous groupings and lends support to grouping strategies that allow LPS students the opportunity to engage with higher-performing students in groups. In our study, these benefits to LPS students were not at the cost of MPS or HPS students. In fact, satisfaction and reported quality of product and process were higher for all students in heterogeneous compared with homogeneous groups.

The success of LPS students in heterogeneous groups might be explained by socially oriented theories of development, which posit that the social environment of a student can initiate and influence change in that student. For example, Vygotsky (1978) suggested that students have a "zone of proximal development" in which they have the capacity to perform at a higher level than their current level of development, and performance could be influenced by the student's academic peers. In our case, LPS students likely needed the knowledge base of their higher-performing peers, resulting in better performance on the postassessment and higher learning gains. This suggests that the low learning gains in low-homogeneous groups may have had more to do with a lack of a knowledge base or necessary work behaviors (work behaviors were not captured on the SAGE survey), rather than group process. Overall, providing students with the opportunity to discuss concepts with their classmates is important in undergraduate classrooms (Mazur, 1996; Tanner, 2009); for LPS students in our study, there is evidence that improved performance was correlated with times when those discussions were with higher-performing students.

There is a common concern among instructors that high-performing students will be frustrated with lower-performing students in their group (Cooper, 1995). However, two of the four results from our SAGE survey directly counter this concern, because students were less frustrated on average with group work and perceived their work as higher quality when they worked in heterogeneous groups compared with when they were placed in homogeneous groups, when controlling for student performance on the preassessment. In addition, Gaudet *et al.* (2010) found that high-performing students were positively affected by having lower-performing students in their group and experienced positive shifts in attitude across all four SAGE factors. All students generally felt positive about group work in our three classes. This was true even for students in all-LPS groups, in which we would expect greater frustration as students received feedback from assessments.

Using Demographic Variables to Assign Groups versus Allowing Students to Self-Select Their Groups

Because LPS students were most successful when working with higher-competency students, we sought to replicate heteroge-

neous group formation by using demographic variables instead of a time-consuming in-class pretest. We chose to use self-reported GPA and number of science classes taken at the university level to assign groups, because we thought these variables would best predict preassessment score and allow us to form heterogeneous competence groups (Freeman *et al.*, 2007; Connell *et al.*, 2016). While these two variables did not capture a large amount of the variance among student preassessment scores, they proved successful for creating groups that were heterogeneous, although many of the groups that formed were composed of students from only two of the three competency levels. However, after we administered the preassessment and compared scores with all the variables on the survey, we found that only self-reported GPA and self-reported proficiency in biology significantly predicted preassessment score and were included in the best-fit models. In some ways, it was surprising that student self-reported proficiency in biology was a significant predictor of competency. Many students struggle to accurately assess their own performance on tests, and low-aptitude students are particularly prone to this miscalculation (Kruger and Dunning, 1999; Dunning *et al.*, 2003). In our classes, students came with some self-awareness about their competency in biology. Thus, it appears that self-reported GPA can be used to create heterogeneous competence groups, and there are likely other variables that could be used in conjunction with self-reported GPA. Presently, our students take a four-question survey (which includes self-reported GPA, self-reported proficiency in biology, gender, and race/ethnicity) on our learning management system before the start of class, and we use these variables to form groups.

In the class in which students could self-select their own groups, we were surprised to find that the resulting groups were heterogeneous and mostly indistinguishable from the distribution of group types in the Demographic class. We had hypothesized that the Self-selected class would have more homogeneous groups than expected by chance due to a "back of the room effect," wherein less confident, lower-performing students could congregate and form groups on the first day of class, resulting in low-homogeneous groups. However, even though the group types were the same, we found lower overall learning outcomes in the Self-selected class compared with the Demographic class, for which the instructor assigned groups. One might hypothesize that the lower learning outcomes in the Self-selected class could be a product of a higher number of groups with all LPS students, but when students were allowed to self-select into groups, no homogeneous LPS student groups formed. Some other process, such as avoiding potentially unproductive friend groups, may explain the higher performance in the class in which the instructor assigned groups. This difference raises a question of whether it was the act of an instructor assigning groups or some other nature of the class dynamics that resulted in the significant difference observed in student learning between these classes. Previous work suggests that the difference observed may indeed be related to the way groups were formed. Although Feichtner and Davis (1984) did not assess group effectiveness by learning outcomes, they did find that students reported more negative experiences in self-selected groups. Colbeck *et al.* (2000) also found that students prefer when faculty intentionally form groups. However, our result must be treated with caution. While both classes were

taught in the same manner by the same instructor, materials and assessments were identical between the two classes, and we controlled for student competence in our analysis, it is not entirely possible to determine whether the observed difference in learning gains stems specifically from how groups were formed or who the students were in the two classes.

Considerations about Forming Groups

Among the three classes in our study, assigning students to groups using demographic variables was the most effective method we tested in terms of instructor time, effort, and student performance. Using data from a survey was less time consuming than an in-class pretest, it did not produce homogeneous low-competence groups, and it yielded higher learning outcomes compared with the class in which students were allowed to self-select. While using demographic data to form groups requires obtaining information from students or the registrar, this proved considerably easier than creating, administering, and grading an in-class pretest. Forming groups using data from the four-question survey that we administer before the start of class and posting student group numbers to our learning management system takes approximately 1 hour of instructor work for our 200-person lecture class.

Although allowing students to self-select their groups takes the least amount of instructor effort, there are foreseeable risks associated with this method. Besides the lower learning outcomes we observed, there remains a risk that that homogeneous groups of low-competence students will form. We did not have any of these groups form in our class, but we only tested this method in one section of Biology 101. Freeman *et al.* (2017) found that when students self-sorted into groups in a large-enrollment biology class, students who had a history of low academic achievement tended to group together, while students who were doing well in the class were more likely to work together as the class progressed, suggesting that students might form homogeneous groups when allowed to choose their group mates. Those authors analyzed pairs of students, however, so their result might not predict what would happen when students are working in larger groups, where this risk is lower. Another risk of allowing students to self-select is heightened anxiety among students who cannot find a group. In our study, students who could not find a group came to the front of the class, where they were assisted by the instructor or teaching assistant. A handful of students expressed anxiety about this process. Predetermining groups reduces the chaos and likely lowers the anxiety for some students.

Other Contexts

Perhaps one of the reasons that there has been conflicting evidence about group formation practices is that the results are context dependent. For nonmajors biology students in our flipped, student-centered classroom, most of whom were traditional students, heterogeneous groupings yielded better leaning outcomes for LPS students and did not help or hinder MPS or HPS students compared with homogeneous groupings. In our cooperative groups, many of the assignments produced group grades (20% of the total class grade was based on in-class group assignments and tests, although these were not part of our analyses), so group members succeeded or failed together. The class was also a low-stakes general education requirement for ~80% of the students,

so the final grade did not contribute to major requirements or entrance into a program. Our results could differ in a more competitive environment in which students are competing for grades or the class is a prerequisite for a highly restricted major or program. Additionally, student demographics could also play a role. Classrooms with a large proportion of nontraditional or returning students might benefit from different types of groupings, especially if there is a substantial amount of group work outside class and nonschool schedules need to be accommodated. Furthermore, the physical space of the classroom may also factor in. This study was conducted in a traditional lecture hall with fixed seating, and perhaps results would differ in a SCALE-UP classroom (Beichner *et al.*, 2007) or another environment more conducive to active learning. Finally, the way instructors frame group work though talk and immediacy may influence the effectiveness of group work and may contribute to the failure or success of group learning (Seidel *et al.*, 2015).

CONCLUSIONS

Leveraging student interactions to maximize cooperative learning experiences in a postsecondary science classroom involved intentionally assigning students into formal heterogeneous groups by content knowledge and allowing them to self-select into heterogeneous groups. Heterogeneous group configurations were associated with greater learning gains for students with the lowest competency coming into the class compared with homogeneous groupings, with minimal impacts between group configurations for other students. In our context, using a short preassessment, two variables from a demographic survey, or allowing students to self-select their own groups, produced heterogeneous student groups. Students in the Self-selected class had lower learning gains compared with the demographically formed class, suggesting a possible underlying mechanism, but further research would be needed to understand whether this finding would hold in other settings or across more classrooms. Given our experiences and investigation of student outcomes, we recommend intentionally forming heterogeneous groups using GPA and student-perceived competence with class content, while balancing for race and gender, to intentionally structure heterogeneous groups.

REFERENCES

Anderson, D. R., & Burnham, K. P. (2002). Avoiding pitfalls when using information-theoretic methods. *Journal of Wildlife Management*, *66*, 912–918.

Bacon, D. R., Stewart, K. A., & Silver, W. S. (1999). Lessons from the best and worst student team experiences: How a teacher can make the difference. *Journal of Management Education*, *23*, 467–488.

Bacon, D. R., Stewart, K. A., & Stewart-Belle, S. (1998). Exploring predictors of student team project performance. *Journal of Marketing Education*, *20*, 63–71.

Baer, J. (2003). Grouping and achievement in cooperative learning. *College Teacher*, *51*, 169–174.

Barton, K. (2009). MuMIn: Multi-model inference, R package (Version 0.12). Retrieved June 10, 2018, from https://r-forge.r-project.org/projects/mumin

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Beichner, R. J., Saul, J. M., Abbott, D. S., Morse, J. J., Deardoff, D., Allain, R. J., … Risley, J. S. (2007). Student-centered activities for large enrollment undergraduate programs (SCALE-UP) project. In Redish, E., & Cooney, P. (Eds.), *Research-based reform of university physics* (pp. 1–42). College Park, MD: American Association of Physics Teachers.

Brickell, J. L., Porter, D. B., Reynolds, M. F., & Cosgrove, R. D. (1994). Assigning students to groups for engineering design projects: A comparison of five different methods. *Journal of Engineering Education*, *83*, 259–262.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretical approach*. New York: Springer.

Colbeck, C. L., Campbell, S. E., & Bjorklund, S. A. (2000). Grouping in the dark: What college students learn from group projects. *Journal of Higher Education*, *71*, 60–83.

Connell, G. L., Donovan, D. A., & Chambers, T. G. (2016). Increasing the use of student-centered pedagogies from moderate to high improves student learning and attitudes about biology. *CBE—Life Sciences Education*, *15*, ar3.

Cooper, M. M. (1995). Cooperative learning: An approach for large enrollment courses. *Journal of Chemical Education*, *72*, 162–164.

Corwin, L. A., Graham, M. J., & Dolan, E. L. (2015). Modeling course-based undergraduate research experiences: An agenda for future research and evaluation. *CBE—Life Sciences Education*, *14*, es1.

D'Avanzo, C., Anderson, C. W., Griffith, A., & Merrill, J. (2010). *Thinking Like a Biologist: Using Diagnostic Questions to Help Students Reason with Biological Principles*. Retrieved January 17, 2010, from www.biodqc.org

Dolan, E. L., Lally, D. J., Brooks, E., & Tax, F. E. (2008). Prepping students for authentic science. *Science Teacher*, *75*, 38–43.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*, 83–87.

Feichtner, S. B., & Davis, E. A. (1984). Why some groups fail: A survey of students' experiences with learning groups. *Journal of Management Education*, *9*, 58–73.

Fischer, K. M., Williams, K. S., & Lineback, J. E. (2011). Osmosis and diffusion conceptual assessment. *CBE—Life Sciences Education*, *10*, 418–429.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences USA*, *111*, 8410–8415.

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., … Wenderoth, M. P. (2007). Prescribed active learning increases performance in introductory biology. *CBE—Life Sciences Education*, *6*, 132–139.

Freeman, S., Theobald, R., Crowe, A. J., & Wenderoth, M. P. (2017). Likes attract: Students self-sort in a classroom by gender, demography, and academic characteristics. *Active Learning in Higher Education*, *18*, 115–126.

Gaudet, A. D., Ramer, L. M., Nakonechny, J., Cragg, J. J., & Ramer, M. S. (2010). Small-group learning in an upper-level university biology class enhances academic performance and student attitudes toward group work. *PLoS ONE*, *5*(12), e15821. doi: 10.1371/journal.pone.0015821

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.) Somerset, NJ: Wiley.

Handelsman, J., Ebert-May, D., Beichner, R., Bruns, P., Chang, A., DeHaan, R., … Wood, W. B. (2004). Scientific teaching. *Science*, *304*, 521–522.

Harlow, J. J. B., Harrison, D. M., & Meyertholen, A. (2016). Effective student teams for collaborative learning in an introductory university physics course. *Physical Review Physics Education Research.*, *12*, 010138. https://doi.org/10.1103/PhysRevPhysEducRes.12.010138

Heller, P., & Hollabaugh, M. (1991). Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups. *American Journal of Physics*, *60*, 637–644.

Jensen, J. L., & Lawson, A. (2011). Effects of collaborative group composition and inquiry instruction on reasoning gains and achievement in undergraduate biology. *CBE—Life Sciences Education*, *10*, 64–73.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher*, *38*, 365–379.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1991). *Active learning: Cooperation in the college classroom*. Edina, MN: Interaction Book Company.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (2014). Cooperative learning: Improving university instruction by basing practice on validated theory. *Journal of Excellence in College Teaching*, *25*, 85–118.

Klymkowsky, M. W., & Garvin-Doxas, K. (2008). Recognizing students' misconceptions through Ed's Tools and the Biology Concept Inventory. *PLoS Biology*, *6*, e3. doi: 10.1371/journal.pbio.0060003

Kouros, C., & Abrami, P. C. (2006) How do students really feel about working in small groups? The role of student attitudes and behaviours in cooperative classroom settings. *Paper presented at: Annual Meeting of the American Educational Research Association, San Francisco, CA*.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessment. *Journal of Personality and Social Psychology*, *77*, 1121–1134.

Kulik, C. L. C., & Kulik, J. (1982). Effects of ability grouping on secondary school students. A meta-analysis of evaluation findings. *American Education Research Journal*, *19*, 415–428.

Kulik, C. L. C., & Kulik, J. (1984). Effects of accelerated instruction on students. *Review of Educational Research*, *54*, 409–425.

Lawrenz, F., & Munch, T. W. (1984). The effect of grouping laboratory students on selected educational outcomes. *Journal of Research in Science Teaching*, *21*, 699–708.

Lou, Y., Abrami, P. C., & Spence, J. C. (2000). Effects of within-class grouping on student achievement: An exploratory model. *Journal of Educational Research*, *94*, 101–112.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollnia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, *66*, 423–458.

Mazur, E. (1996). *Peer instruction: A user's manual*. Englewood Cliffs, NJ: Prentice-Hall.

McInerney, M. J., & Fink, L. D. (2003). Team-based learning enhances long-term retention and critical thinking in an undergraduate microbial physiology course. *Microbiology Education*, *4*, 3–12.

Mello, J. A. (1993). Improving individual member accountability in small work group settings. *Journal of Management Education*, *17*, 253–259.

Michaelsen, L. K., Knight, A. B., & Fink, L. D. (2004). *Team-based learning: A transformative use of small groups in college teaching*. Sterling, VA: Stylus.

Michaelsen, L. K., & Sweet, M. (2008). The essential elements of team-based learning. *New Directions for Teaching and Learning*, *116*, 7–27.

Miller, H. B., Witherow, D. S., & Carson, S. (2012). Student learning outcomes and attitudes when biotechnology lab partners are of different academic levels. *CBE—Life Sciences Education*, *11*, 323–332.

Nadelson, L. S., & Southerland, S. A. (2010). Development and preliminary evaluation of the measure of understanding of macroevolution: Introducing the MUM. *Journal of Experimental Education*, *78*, 151–190.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rosser, S. V. (1998). Group work in science, engineering, and mathematics: Consequences of ignoring gender and race. *College Teacher*, *46*, 82–88.

Ruiz-Primo, M. A., Briggs, D., Iverson, H., Talbot, R., & Shephard, L. A. (2011). Impact of undergraduate science course innovations on learning. *Science*, *331*, 1269–1270.

Seidel, S. B., Reggi, A. L., Schindler, J. N., Burrus, L. W., & Tanner, K. D. (2015). Beyond the biology: A systematic investigation of noncontent instructor talk in an introductory biology course. *CBE—Life Sciences Education*, *14*, ar43.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, *60*, 471–499.

Smith, B. L., & McGregor, J. T. (1992). What is collaborative learning?" In Goodsell, A. S., Maher, M. R., & Tinto, V. (Eds.), *Collaborative learning: A*

sourcebook for higher education. Syracuse, NY: National Center on Postsecondary Teaching, Learning, & Assessment, Syracuse University.

Smith, M. K., Wood, W. B., Krauter, K., & Knight, J. K. (2011). Combining peer discussion with instructor explanation increases student learning from in-class concept questions. *CBE—Life Sciences Education*, *10*, 55–63.

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research*, *69*, 21–51.

Strong, J. T., & Anderson, R. E. (1990). Free-riding in group projects: Control mechanisms and preliminary data. *Journal of Marketing Research*, *12*, 61–67.

Tanner, K. (2009). Talking to learn: Why biology students should be talking in classrooms and how to make it happen. *CBE—Life Sciences Education*, *8*, 89–94.

Tanner, K., Chatman, L. S., & Allen, D. (2003). Approaches to cell biology teaching: Cooperative learning in the science classroom—Beyond students working in groups. *Cell Biology Education*, *2*, 1–5.

Theobald, E. J. (2018). Students are rarely independent: When, why, and how to use random effects in discipline-based education research. *CBE—Life Sciences Education*, *17*, rm2.

Theobald, E. J., Eddy, S. L., Grunspan, D. Z., Wiggins, B. L., & Crowe, A. J. (2017). Student perception of group dynamics predicts individual performance: Comfort and equity matter. *PLoS ONE*, *12*, e0181336.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Watson, S. B., & Marshall, J. E. (1995). Effects of cooperative incentives and heterogeneous arrangement on achievement and interaction of cooperative learning groups in a college life science course. *Journal of Research in Science Teaching*, *32*, 291–299.